Genome diversification in Staphylococcus aureus

INFECTION AND IMMUNITY, May 2003, p. 2827–2838 Vol. 71, No. 5
0019-9567/03/$08.00+0 DOI: 10.1128/IAI.71.5.2827–2838.2003

Genome Diversification in *Staphylococcus aureus*: Molecular Evolution of a Highly Variable Chromosomal Region Encoding the Staphylococcal Exotoxin-Like Family of Proteins

J. Ross Fitzgerald,[1]† Sean D. Reid,[1] Eeva Ruotsalainen,[2] Timothy J. Tripp,[3] MengYao Liu,[1] Robert Cole,[1] Pentti Kuusela,[4] Patrick M. Schlievert,[3] Asko Järvinen,[2] and James M. Musser[1]*

Laboratory of Human Bacterial Pathogenesis, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Hamilton, Montana 59840[1]; Department of Medicine, Division of Infectious Diseases,[2] and Division of Clinical Microbiology, HUCH Laboratory Diagnostics,[4] Helsinki University Central Hospital, Finland; and Department of Microbiology, University of Minnesota Medical School, Minneapolis, Minnesota 55455[3]

Received 20 November 2002/Returned for modification 14 January 2003/Accepted 12 February 2003

Recent genomic studies have revealed extensive variation in natural populations of many pathogenic bacteria. However, the evolutionary processes which contribute to much of this variation remain unclear. A previous whole-genome DNA microarray study identified variation at a large chromosomal region (RD13) of *Staphylococcus aureus* which encodes a family of proteins with homology to staphylococcal and streptococcal superantigens, designated staphylococcal exotoxin-like (SET) proteins. In the present study, RD13 was found in all 63 *S. aureus* isolates of divergent clonal, geographic, and disease origins but contained a high level of variation in gene content in different strains. A central variable region which contained from 6 to 10 different *set* genes, depending on the strain, was identified, and DNA sequence analysis suggests that horizontal gene transfer and recombination have contributed to the diversification of RD13. Phylogenetic analysis based on the RD13 DNA sequence of 18 strains suggested that loss of various *set* genes has occurred independently several times, in separate lineages of pathogenic *S. aureus*, providing a model to explain the molecular variation of RD13 in extant strains. In spite of multiple episodes of *set* deletion, analysis of the ratio of silent substitutions in *set* genes to amino acid replacements in their products suggests that purifying selection (selective constraint) is acting to maintain SET function. Further, concurrent transcription in vitro of six of the seven *set* genes in strain COL was detected, indicating that the expression of *set* genes has been maintained in contemporary strains, and Western immunoblot analysis indicated that multiple SET proteins are expressed during the course of human infections. Overall, we have shown that the chromosomal region RD13 has diversified extensively through episodes of gene deletion and recombination. The coexpression of many *set* genes and the production of multiple SET proteins during human infection suggests an important role in host-pathogen interactions.

Staphylococcus aureus causes a variety of diseases in humans and animals and produces a large number of secreted proteins which contribute to infection (4, 7). Recent comparative genomic studies have revealed a high level of interstrain variation in genome content, particularly at regions containing genes encoding virulence factors or antibiotic resistance mechanisms (2, 5, 6). For example, a recent DNA microarray study examining genomic variation in *S. aureus* identified 18 large chromosomal regions of difference among 36 strains isolated from different infection types in humans, cows, and sheep (6). Among the 18 large chromosomal regions identified, one (RD13) varied in size from 12 to 17 kb and was predicted to contain considerable variation in gene content (Fig. 1). RD13 corresponds to the exotoxin gene-containing regions of genomic islands SaPIn2 and SaPIm2, identified in *S. aureus* strains N315 and Mu50, respectively (10). RD13 has open reading frames encoding hypothetical proteins, a transposase, a restriction-modification system, and at least seven staphylococcal exotoxin-like (SET) proteins, depending on the strain (Fig. 1). Analysis of the inferred amino acid sequences indicates that SET proteins contain internal regions of homology with superantigens made by *S. aureus* and group A *Streptococcus*. Superantigens produced by *S. aureus* and group A *Streptococcus* are thought to modulate the host immune response during infection by binding and activating T-cell subsets expressing specific Vβ chains of the T-cell receptor (4, 11, 14). However, a very recent study reported that a purified recombinant SET variant did not have the classical characteristics of superantigens, such as mitogenicity, pyrogenicity, or the enhancement of endotoxic shock (1). A representative SET (SET3) has the classical three-dimensional structure which is characteristic of superantigens, but there are some notable differences perhaps reflecting an alternative function. In particular, SET3 has a large, positively charged, saddle-shaped surface that has the potential to act as

* Corresponding author. Mailing address: Laboratory of Human Bacterial Pathogenesis, Rocky Mountain Laboratories, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 903 South 4th St., Hamilton, MT 59840. Phone: (406) 363-9315. Fax: (406) 363-9427. E-mail: jmusser@niaid.nih.gov.

† Present address: Department of Microbiology, Moyne Institute of Preventive Medicine, University of Dublin, Trinity College, Dublin 2, Ireland.

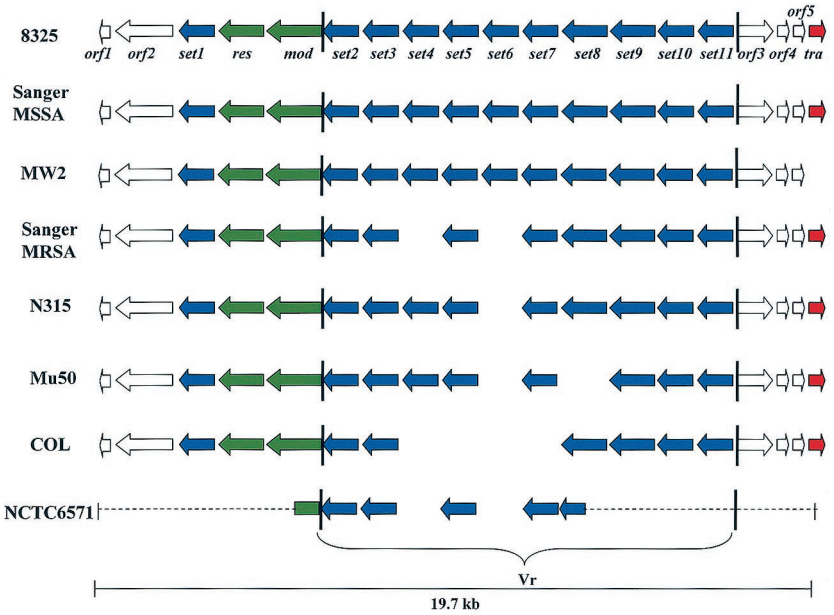

FIG. 1. Genetic structure of chromosomal region RD13 in eight *S. aureus* strains: 8325, Sanger MSSA, MW2, Sanger MRSA, N315, Mu50, COL, and NCTC6571. The proteins encoded by the genes designated are as follows: *orf1* to *orf5* (white), hypothetical proteins; *set1* to *set11* (blue), staphylococcal exotoxin-like proteins; *res* and *mod* (green), restriction-modification subunits; *tra* (red), transposase. Dashed lines represent DNA of unknown sequence. The central Vr is indicated.

a binding surface for negatively charged molecules such as DNA (1). However, the function of this family of proteins has yet to be established.

The precise molecular processes leading to genome diversification in pathogenic bacteria remain unclear. Many of the variable chromosomal regions identified in the genome of *S. aureus* are related to insertion elements, transposons, phage, and pathogenicity islands and are found in only a portion of strains examined, indicating the role of horizontal gene transfer in *S. aureus* evolution. However, a large region (RD13) of the chromosome with extensive variation in virulence gene content was identified in all strains examined by microarray analysis. The unusual variation in nucleotide and gene content of RD13 raised important questions regarding the processes that have contributed to the genetic diversity of RD13 and to the evolution of the *S. aureus* genome in general. In the present study, comparative sequencing, PCR-restriction fragment length polymorphism (RFLP) analysis, gene and protein expression assays, and evolutionary genetic analyses were used to investigate the molecular evolution and genetic diversity of RD13 in strains of *S. aureus* isolated from human and animal infections. A model for the evolutionary history of RD13 among pathogenic clones is proposed.

MATERIALS AND METHODS

***S. aureus* strains.** Strains were selected to represent the most abundant clonal lineages identified in a multilocus enzyme electrophoresis (MLEE) population genetic study of over 2,000 *S. aureus* isolates and were of broad geographic and disease origin (Table 1) (12).

RD13 PCR-RFLP analysis. PCR primers (forward primer, 5′-CAA AGC TAA ACA AGA CGG CTT TGA TG-3′; reverse primer, 5′-TCC GCG CCA ATC TTC TGG AAC-3′) specific for conserved regions flanking RD13 were used to amplify the intervening sequence with the Advantage genomic PCR kit (Clontech, Palo Alto, Calif.) according to the manufacturer's instructions. The PCR products were digested with restriction endonuclease *Hin*P1I at 37°C for 2 h, and the DNA restriction fragments were resolved by electrophoresis in a 1.2% agarose gel.

Analysis of size variation in Vr in RD13. The size of the RD13 variable region (Vr) in each strain was inferred on the basis of amplification by PCR with primers (forward primer, 5′-ATA GAA CTC GCC TGC TTT TTT ACC-3′; reverse primer, 5′-CCA AAG CCT TTA GGT TCA TCA TAC A-3′) specific for conserved flanking regions (the *mod* and *orf3* genes, respectively) (Fig. 1)

DNA sequence analysis. Sequence data obtained from both DNA strands with an Applied Biosystems model 3700 automated sequencer were analyzed by DNASTAR (Madison, Wis.). Multiple-sequence alignment of the inferred amino acid sequences was performed with Clustal W version 1.8 (18). Statistics of nucleotide and amino acid content were determined with MEGA version 2.1 (9).

Population and molecular evolutionary genetic analyses. The Nei-Gojobori method (13) was used to calculate the proportion of synonymous sites in nucleotide sequence data. Phylogenies were constructed with the neighbor-joining algorithm by using MEGA version 2.1 (9). A concatenated sequence consisting of *orf2*, *set1*, *res*, *mod*, *set2*, *set3*, *set9*, *set10*, and *set11* was used for the analysis of the conserved architecture of RD13. Due to strain-to-strain variation in the content of certain *set* genes, *set4*, *set5*, *set6*, *set7*, and *set8* were studied independently. To examine variation in the patterns of nucleotide substitution, the proportion of synonymous (p_S) and the proportion of nonsynonymous (p_N) nucleotide substitutions were calculated by sliding-window analysis of 30 codons along each gene with the program PSWIN (15) (PSWIN is available from S. D. Reid [sreid@niaid.nih.gov]). Estimates of the sampling variance of these statistics were made by Monte Carlo simulation or by bootstrapping.

To identify the putative end points of past recombination events, a computer program (MAXCHI) that implements the maximum chi-square method was used (16). Happlot analysis was used to diagram the locations of polymorphic sites.

Real-time reverse transcriptase PCR (TaqMan assays). PCR primers and probes were designed with the software package Primer Express (Perkin-Elmer) on the basis of the genomic DNA sequence available for *S. aureus* strain COL (http://www.tigr.org) and purchased from PE Applied Biosytems. Total RNA was isolated from bacteria cultured for 3 h (mid-exponential phase) or 12 h (stationary phase), and reverse transcription and PCRs were performed as described previously (3). To confirm that the *era* gene encoding the GTP-binding protein Era was constitutively expressed and suitable for use as an internal control, the amount of *era* mRNA was measured in three separate reverse transcriptase PCR experiments with two independent RNA samples isolated from mid-exponential

TABLE 1. PCR-RFLP types of *S. aureus* strains of divergent clonal lineages

Strain[a]	Lineage/electrophoretic type[b]	Disease or characteristic[c]	Source (date of isolation)[d]	PCR-RFLP type	Vr size (kb)
COL	ND	MR	Colindale, United Kingdom	1	7
MSA 1401	A1/1	Bovine mastitis	Cornell	25	9
MSA 951	A1/5	Bovine mastitis	Louisiana	24	10
MSA 948	A1/7	Bovine mastitis	Puerto Rico	22	11
MSA 2050	B1/10	Endocarditis	Denmark (1985)	23	9
MSA 2099	D2/32	Endocarditis	Denmark (1984)	12	9
MSA 2389	D2/39	Furunculosis	Sweden	12	9
MSA 2967	D3/45	Sepsis	Canada	13	9
MSA 1601	D3/53	MR	Rhode Island	13	9
MSA 915	E1/61	Bovine mastitis	Louisiana	18	10
MSA 535	F1/66	Ovine mastitis	Germany	11	11
MSA 551	F1/70	Ovine mastitis	France	11	11
MSA 3402	F2/89	MR	New York (1978)	16	11
MSA 3426	F2/89	MR	Dublin, Ireland (1980s)	3	11
MSA 820	F2/91	MR	Rhode Island	6	10
MSA 890	F2/93	MR	Texas	2	10
MSA 3400	F2/91	MR	Dublin, Ireland (1990)	6	10
MSA 3405	F2/91	MR	Canada (1980s)	7	9
MSA 3410	F2/93	MR	London, United Kingdom (1960s)	1	7
MSA 1006	F3/91	Bovine mastitis	Louisiana	3	11
MSA 1007	F3/106	Bovine mastitis	Louisiana	3	11
MSA 817	F4/114	Human origin	Rhode Island (1980s)	4	10
MSA 961	F4/146	Bovine mastitis	Louisiana	5	10
MSA 2120	F4/146	Endocarditis	Denmark (1983)	8	11
MSA 2766	F5/161	TSS	Canada	32	9
MSA 1605	F5/164	Turkey hock	Utah	19	11
MSA 1260	F5/165	Vaginal commensal	P. Schlievert	6	10
MSA 573	F6/170	TSS	B. Kreiswirth	20	11
MSA 581	F6/170	Human origin	B. Kreiswirth	19	11
MSA 565	F8/178	Human origin	B. Kreiswirth	3	9
MSA 2020	F9/189	Scalded-skin syndrome	France	10	11
MSA 2965	F9/191	Sepsis	Canada (1983)	10	11
MSA 2348	F9/189	Furunculosis	Sweden	10	11
RF122	F10/195	Bovine mastitis	Ireland (1993)	9	9
MSA 632	F11/205	TSST-1$^+$	B. Kreiswirth	21	11
MSA 1184	F11/205		W. Karakawa	21	11
MSA 554	GI/213	Chicken	B. Kreiswirth	17	8
MSA 537	H1/234	TSS	United States (1985)	15	9
MSA 2335	H1/234	TSS	Sweden	15	10
MSA 2885	H1/234	Nasal commensal	Canada (1974)	15	10
MSA 3407	H1/234	TSS	New York (1978)	15	10
MSA 3412	H1/234	TSS	New York (1980s)	14	7
MSA 1183	I1/245		R. Proctor	32	9
MSA 1134	K1/249	Human TSST-1$^+$	P. Schlievert	6	10
MuSA 141	ND	Septicemia (w/o)	Finland (1999)	33	9
MuSA 142	ND	Septicemia (w)	Finland (1999)	16	9
MuSA 143	ND	Septicemia (w/o)	Finland (1999)	13	9
MuSA 144	ND	Septicemia (w)	Finland (1999)	6	10
MuSA 145	ND	Septicemia (w/o)	Finland (1999)	26	10
MuSA 146	ND	Septicemia (w)	Finland (1999)	8	11
MuSA 147	ND	Septicemia (w/o)	Finland (1999)	34	8
MuSA 148	ND	Septicemia (w)	Finland (1999)	27	11
MuSA 149	ND	Septicemia (w/o)	Finland (1999)	13	9
MuSA 150	ND	Septicemia (w)	Finland (1999)	28	10
MuSA 231	ND	Septicemia (w/o)	Finland (2000)	13	9
MuSA 232	ND	Septicemia (w)	Finland (2000)	34	8
MuSA 233	ND	Septicemia (w)	Finland (2000)	—[e]	10
MuSA 234	ND	Septicemia (w/o)	Finland (2000)	13	9
MuSA 236	ND	Septicemia (w/o)	Finland (2000)	30	9
MuSA 237	ND	Septicemia (w)	Finland (2000)	13	9
MuSA 238	ND	Septicemia (w)	Finland (2000)	15	9
MuSA 239	ND	Septicemia (w)	Finland (2000)	31	10
MuSA 240	ND	Septicemia (w/o)	Finland (2000)	34	8

[a] MSA and MuSA, Musser *S. aureus* strain designations.
[b] Phylogenetic lineage and electrophoretic type as designated by Musser and Selander (12). ND, not determined.
[c] MR, methicillin resistant; TSS, toxic shock syndrome; TSST-1$^+$, toxic shock syndrome toxin-1 positive; w, with deep infection foci; w/o, without deep infection foci.
[d] P. Schlievert, culture collection of P. Schlievert; B. Kreiswirth, culture collection of B. Kreiswirth; W. Karakawa, culture collection of W. Karakawa; R. Proctor, culture collection of R. Proctor.
[e] —, no product.

TABLE 2. Primers, vectors, cloning sites, and recombinant plasmids used in gene cloning

Gene	Strain	Primers	Vector	Cloning sites	Plasmid
set1	COL	5′-CCAGTACATATGAGTACATTAGAGGTTAGATCA-3′, 5′-CGGATCCTAATATAAATCGACTTCAATTT-3′	pET21b	*Nde*I, *Bam*HI	*pset*2289
set2	COL	5′-CAGGTCATATGAAACAAAATCAAAAGTCAGTA-3′, 5′-AGGATCCTACTTTAAGTTAACTTCAATATC-3′	pET21b	*Nde*I, *Bam*HI	*pset*2293
set3	COL	5′-CTGGTCATATGAAAGTAGAACTTGATGAGACA-3′, 5′-CGGATCCTAATTCAAATTCACTTCAATAT-3′	pET21b	*Nde*I, *Bam*HI	*pset*2294
set4	8325	5′-CGGTTCATATGAAAGGAAAGTATGAAAAAATG-3′, 5′-CGGATCCTATTTCAAATTCACTTCGATGT-3′	pET21b	*Nde*I, *Bam*HI	*pset*8325-1
set5	8325	5′-GCAGTTCATATGAAAGAAAAGCAAGAGAGAGTA-3′, 5′-CGGATCCTACTTACTTTAAATTTGTTTCA-3′	pET21b	*Nde*I, *Bam*HI	*pset*8325-2
set6	8325	5′-CAGTACATATGGCAGAATCAACTCAAGGTCAA-3′, 5′-CGGATCCTATTTATATTCTAGCTCAACAT-3′	pET21b	*Nde*I, *Bam*HI	*pset*8325-3
set7	8325	5′-CTGTAACCATGGGTGAACATAAAGCAAAATAT-3′, 5′-AGGATCCTATCTAATGTTGGCTTCTATTTT-3′	pET15	*Nco*I, *Bam*HI	*pset*8325-4
set8	COL	5′-GCAGCACATATGACAACACCATCTTCCACTAAA-3′, 5′-CGGATCCTATTTCAAATTCACTTCGATGT-3′	pET21b	*Nde*I, *Bam*HI	*pset*2295
set9	COL	5′-CGGTCCATATGGAAAAAATACAATCAACT-3′, 5′-TGGATCCTATTTTATATTCACTTCAATTG-3′	pET21b	*Nde*I, *Bam*HI	*pset*2298
set10	COL	5′-CCAGTTCATATGGAAAAGAAACCTATTGTAATA-3′, 5′-TGGATCCTATGCTTTTATAACTTTGATTTG-3′	pET21b	*Nde*I, *Bam*HI	*pset*2300
set11	COL	5′-CAGTACATATGAAAGCAGAAGTTAAACAACAA-3′, 5′-AGGATCCTATTTCATTTCTACTAGAATTT-3′	pET21b	*Nde*I, *Bam*HI	*pset*2301

and stationary-phase cultures of strain COL. The results confirmed that *era* is expressed at similar levels (0.9 to 1.3 relative expression levels) in both phases of the growth cycle (data not shown).

***set* gene cloning and production of recombinant SET proteins.** The primers, vectors, and restriction enzymes used for PCR amplification and cloning of the *set* genes of *S. aureus* strains COL and 8325 are listed in Table 2. The primers were designed on the basis of genome sequences available for *S. aureus* strains COL (http://www.tigr.org) and 8325 (http://www.genome.ou.edu), and restriction enzyme sites for cloning were included in each primer. PCR products were amplified with genomic DNA isolated from strain COL or strain 8325, digested with the appropriate restriction enzymes, and cloned into vector pET21b or pET15. The cloned genes were sequenced to ensure that spurious mutations had not been introduced, and the plasmids were moved into *Escherichia coli* BL21 for expression of recombinant proteins.

To assess the production of recombinant proteins, BL21 cells with recombinant plasmids were grown at 37°C for 10 h in 3 ml of Luria-Bertani broth supplemented with 100 mg of ampicillin per liter. Cells were pelleted by centrifugation and suspended in 1× sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) loading buffer at a ratio of 100 μl of buffer per 1 unit of optical density 600 nm per ml. The samples were boiled for 10 min and analyzed by SDS-PAGE.

Purification of recombinant SET protein. To purify recombinant SET6 protein, BL21 cells with recombinant plasmid were grown overnight at 37°C in Luria-Bertani broth. The culture was treated with 4 volumes of absolute ethanol for 48 h, and the precipitated proteins were collected by centrifugation (500 × *g*, 10 min). Five liters of total culture volume was combined in a single culture toxin preparation. The precipitate was air dried after centrifugation and resuspended in 100 to 150 ml of pyrogen-free water. The resuspended precipitate was centrifuged at 10,000 × *g* for 30 min; the concentrated supernatant was removed, placed in dialysis tubing with a 12,000- to 14,000-molecular-weight cutoff (Spectrum Laboratories, Inc., Miami, Fla.), and dialyzed overnight against 4 liters of distilled water. The dialyzed supernatant was subjected to preparative isoelectric focusing. Successive gradients of pH 3.5 to 10 and pH 6 to 8 were used to isolate highly purified recombinant SET6 protein. Final purification of proteins was accomplished with a gel filtration column (Bio-Rad Laboratories, Hercules, Calif.) containing Sephadex G-75 (Sigma, St. Louis, Mo.). Purity was verified by SDS-PAGE in which 10 μg gave a homogenous band of the appropriate molecular weight. The concentration of the purified protein was assessed with the Bradford protein assay (Bio-Rad) (2a), and the protein was lyophilized and stored until used in biological and biochemical assays.

Assay for superantigen activity. Rabbit splenocytes or human peripheral blood mononuclear cells (PBMCs) were seeded into the wells of a 96-well microtiter plate at a concentration of 2×10^5 cells per well. Serial 10-fold dilutions of SET or toxic shock syndrome toxin 1 (TSST-1; positive control) were added to each well in quadruplicate, starting with 1 μg/well and with dilution to 10^{-8} μg/well. The assay results for these dilutions were compared to those for cells incubated in the presence of phosphate-buffered saline alone. The splenocytes were grown at 37°C for 3 days and pulsed with 1 μCi of [^{3}H]thymidine overnight. The cells were harvested the next day, and cell proliferation (incorporation of ^{3}H into DNA) was measured with a scintillation counter (Beckman Instruments, Fullerton, Calif.).

Pyrogenicity and endotoxin enhancement. American Dutch belted rabbits were injected intravenously with recombinant SET proteins at a maximal dose of 10 μg/kg of body weight per ml. The temperature of each rabbit ($n = 3$) was measured with a rectal thermometer at 0 and 4 h postinjection. After 4 h, each rabbit was injected intravenously with 10 μg of lipopolysaccharide from *Salmonella enterica* serovar Typhimurium (1/50 of the 50% lethal dose of endotoxin alone). The lethality of this toxin regimen was assessed over a 48-h period.

Western immunoblot analysis. Proteins separated by SDS-PAGE were transferred to nitrocellulose membranes (Immobilon-NC; Millipore Corporation) with Towbin transfer buffer with a Trans-Blot SD semidry transfer cell (Bio-Rad Laboratories) at 15 V for 40 min. The membrane was incubated with 10 ml of block solution (5% powdered milk in 150 mM NaCl and 100 mM Tris-HCl, pH 7.4) for 1 h, incubated for 1 h with a 1:500 dilution of human patient serum in block solution, rinsed twice, and washed three times for 15 min each time with 0.1% Tween 20 in phosphate-buffered saline. The membrane was incubated with goat anti-human immunoglobulin G horseradish peroxidase-conjugated secondary antibody (Sigma) for 1 h and rinsed and washed as described above. Immunoreactivity was visualized by enhanced chemiluminescence. The human serum samples were obtained during a countrywide study of invasive episodes of *S. aureus* infection in Finland in 1999 and 2000. Patients had infections with or without deep infection foci, including abscesses, cellulitis, pneumonia, and purulent arthritis. Sera were obtained 2 to 7 days (i.e., the acute phase) and 20 to 30 days (i.e., the convalescent phase) after the first positive blood culture. Control serum samples were obtained from four female and four male healthy volunteers, including Caucasian, African-American, Hispanic, and Asian individuals ($n = 2$ each).

Nomenclature of the SET family of proteins. To date in the literature, there has been no coherent system regarding the numbering of the members of the SET protein family. All *S. aureus* strains examined to date, including the genetically divergent strains in the present study, contained between 7 and 11 different *set* genes in RD13. Comparison of RD13 chromosomal regions from different strains indicates that the overall *set* gene order of RD13 is conserved and that *set* genes at the same position in RD13 in different strains are allelic variants of each other (85 to 100% homology). Accordingly, we propose that the *set* gene family of RD13 should be named *set1* to *set11* in consecutive order based on strains with a full complement of *set* genes in RD13 (Fig. 1). To differentiate between allelic variants of *set* genes found in different strains, the gene number should be prefixed by the strain name, e.g., COL*set1*.

TABLE 3. Mean proportion of homologous nucleotide and amino acid sites within six SET genes and proteins common to seven *S. aureus* strains[a]

Gene	Inferred amino acid length in strain COL	Proportion of homologous nucleotide sites[b] (mean ± SE)	Proportion of homologous amino acid sites[c] (mean ± SE)
set1	233	0.81 ± 0.01	0.77 ± 0.02
set2	239	0.95 ± 0.01	0.94 ± 0.01
set3	243	0.95 ± 0.01	0.93 ± 0.01
set9	359	0.92 ± 0.01	0.87 ± 0.01
set10	262	0.95 ± 0.01	0.92 ± 0.01
set11	246	0.89 ± 0.01	0.83 ± 0.02

[a] *S. aureus* strains analyzed were 8325, Sanger MSSA, Sanger MRSA, N315, Mu50, COL, and MW2.
[b] Mean proportion of homologous nucleotide sites = 1 − mean *p*-distance (in nucleotides). *p*-distance is the proportion (*p*) of nucleotide sites at which the sequences compared are different. The standard error was calculated by the bootstrap method.
[c] Mean proportion of homologous amino acid sites = 1 − mean *p*-distance (in amino acids). The *p*-distance is the proportion (p) of amino acid sites at which the sequences compared are different. The standard error was calculated by the bootstrap method.

RESULTS

Characterization of structural variation in RD13. Analysis of the RD13 region in seven sequenced *S. aureus* strains, COL (http://www.tigr.org), 8325 (http://www.genome.ou.edu), methicillin-resistant *S. aureus* (MRSA) 252 (Sanger MRSA) and methicillin-susceptible *S. aureus* (MSSA) 476 (Sanger MSSA) (http://www.sanger.ac.uk), N315 and MW2 (http://www.bio.nite.go.jp), and Mu50 (GenBank accession no. NC_002758), revealed that the gene order at this chromosomal site is well conserved in all isolates. Allelic variants of six of the seven *set* genes contained in RD13 in strain COL were present in all seven sequenced strains (Fig. 1). A summary of the size and degree of polymorphism among *set* genes in the seven sequenced strains is represented in Table 3. Analysis of the degree of polymorphism among predicted SET proteins from a single strain, COL, revealed 22.8 to 64.5% identity, and pairwise analysis of allelic variation among *set* genes common to the seven sequenced strains revealed between 85 and 100% identity at the nucleotide level. In addition, between 6 and 10 *set* variant genes were present in a central Vr depending on the strain (Fig. 1).

In a previous study, Williams et al. (19) sequenced a 5.2-kb DNA segment containing *set* genes from a human *S. aureus* strain, NCTC6571 (Fig. 1). This fragment is identical to a region internal to RD13 in the sequenced MRSA strain 252 and includes *set2*, *set3*, *set5*, *set7*, part of *set8*, and part of the restriction-modification subunit gene. The prototype *set1* gene which was cloned and overexpressed in *E. coli* in the Williams et al. study is an allele of *set5* of the *set* gene family (Fig. 1).

PCR-RFLP analysis of RD13. To analyze structural variation in chromosomal region RD13 among *S. aureus* strains representing the breadth of diversity within the species, a PCR-RFLP method was developed. We analyzed 44 strains that represent each of the major clonal complexes identified in an MLEE population genetics study of more than 2,000 isolates from diverse infection types, localities, and host species (12) (Table 1). Twenty-seven distinct PCR-RFLP types were identified (Fig. 2; Table 1), indicating that substantial variation exists at this chromosomal region in natural populations. Each strain had 7 to 12 restriction fragments in total, which were between 0.5 and 4.5 kb in length, and resulted in restriction profiles that were readily distinguishable (Fig. 2). Most of the 27 PCR-RFLP types comprised relatively few isolates (range,

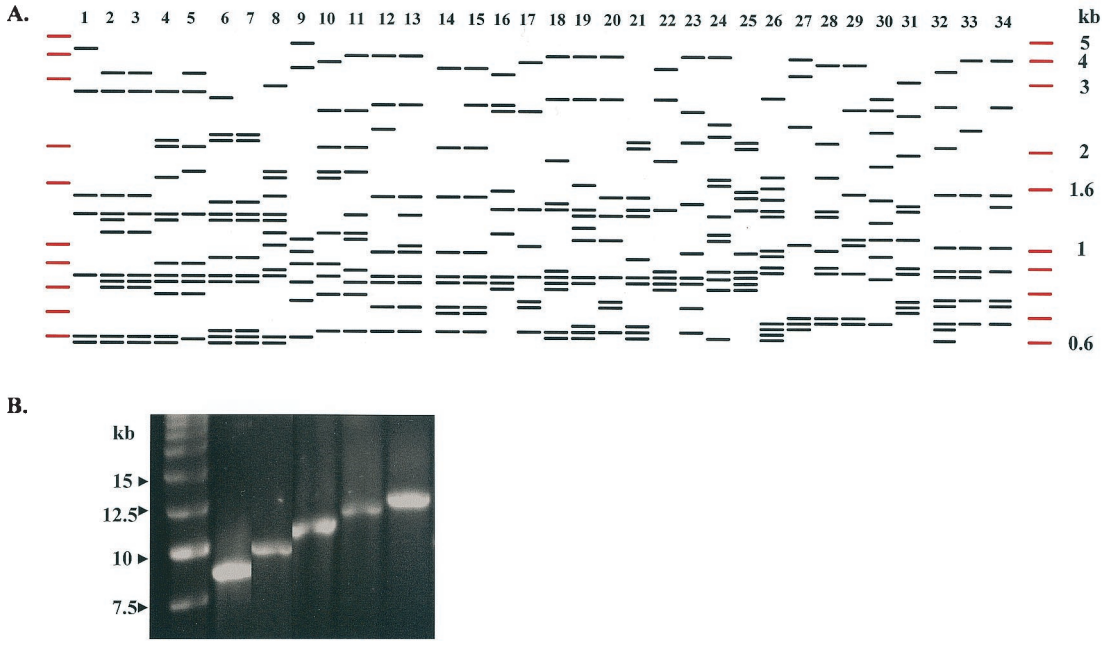

FIG. 2. (A) Schematic representation of 34 RD13 PCR-RFLP types identified among 44 *S. aureus* strains of divergent clonal lineage and 19 strains isolated from patients with invasive disease in Finland; (B) Vr PCR product size variants. Agarose gel electrophoresis of representative PCR products amplified with primers specific for regions flanking the RD13 variable region.

one to six strains per type) (Table 1), which is indicative of considerable genetic diversity in this region. Multiple PCR-RFLP types were identified among isolates assigned to electrophoretic type (ET) clusters A1, F2, F4, F5, F6, and H1. There was little sharing of PCR-RFLP types between ETs. The exceptions to this were strain MSA 565 (lineage F8/ET 178), which shared PCR-RFLP type 3 with strains MSA 3426 (F2/89), MSA 1006 (F3/91), and MSA 1007 (F3/1006); and strain MSA 1134 (K1/249), which had an identical PCR-RFLP type to strains MSA 820 and MSA 3400 (both F2/91). In addition, MSA 2766 (F5/161) had an identical PCR-RFLP type 32 to MSA 1183 of the I1 cluster. This PCR-RFLP typing method had a very high index of discrimination of 0.962 (8). (An index of 1.0 would indicate that a typing method was able to distinguish each member of a strain population from all other members, whereas an index of 0.0 would indicate that all members of the population were of an identical type.) Taken together, these data clearly indicate a high degree of genetic diversity at this chromosomal region.

We also analyzed the RD13 PCR-RFLP patterns of 19 *S. aureus* isolates recovered from patients with invasive infections whose sera were used in Western immunoblot analysis (see below). Among these 19 organisms, 7 had RD13 PCR-RFLP profiles which were identical to type 13 or differed by the presence of one or two restriction sites (types 33 and 34). Two isolates were PCR-RFLP type 16; there were single isolates of type 2, type 6, type 8, and type 15; and five isolates had unique RD13 PCR-RFLP profiles (types 27 to 31). A PCR product could not be generated for one isolate.

PCR analysis of the Vr of RD13. PCR amplification was used to determine the size of the Vr among the 63 strains used in this study (Fig. 2). The primers used were specific for nucleotides located approximately 700 and 800 bp upstream and downstream, respectively, of the Vr. The estimated size of the Vr was calculated by subtracting 1.5 kb from the size of the PCR product. Three strains had Vrs of 7 kb, 5 strains had Vrs of 8 kb, 19 strains had Vrs of 9 kb, 17 strains had Vrs of 10 kb, and 19 strains had Vrs of 11 kb (Table 1). Based on the *set* gene content of the Vrs in the six sequenced strains, these results are consistent with the presence of 6, 7, 8, 9, or 10 *set* genes in this chromosomal region in *S. aureus* strains.

Phylogenetic analysis of RD13. Phylogenetic analysis was used to reconstruct the evolutionary history of the RD13 locus and to determine the extent that recombination contributed to the divergence of the region (Fig. 3). To add to the publicly available genome sequences of *S. aureus* strains 8325, Sanger MSSA, Sanger MRSA, N315, Mu50, and COL, DNA sequencing of variably present *set* genes (Vr) of 12 selected strains was carried out. Strains were selected based on Vr size variation and clonal diversity (determined by MLEE) (strains 1006, 1260, 141, 143, 145, 147, 232, 240, 3400, 3402, 537, and 554). The *set4*, *set5*, *set6*, *set7*, and *set8* genes were found to be variably absent in the strains examined. A phylogenetic tree was constructed from the concatenated nucleotide sequences of *orf2*, *set1*, *res*, *mod*, *set2*, *set3*, *set9*, *set10*, and *set11* (i.e., genes in RD13 that are common to all strains) from the previously sequenced strains 8325, Sanger MSSA, Sanger MRSA, N315, Mu50, and COL to represent the conserved architecture of the RD13 locus. The topology of this tree was compared to the topologies of the individual gene trees of *set4* (strains MSA 1006, MSA 1260, MuSA 145, MSA 3400, MSA 3402, 8325, Sanger MSSA, Mu50, and N315), *set5* (strains MSA 1006, MSA 1260, MuSA 141, MuSA 143, MuSA 145, MSA 3400, MSA 3402, MSA 537, 8325, Sanger MRSA, Sanger MSSA, Mu50, N315, and NCTC6571), *set6* (strains MSA 1006, MSA 3402, 8325, and Sanger MSSA), and *set7* (strains MSA 1006, MSA 1260, MuSA 141, MuSA 143, MuSA 145, MuSA 147, MuSA 232, MuSA 240, MSA 3400, MSA 3402, MSA 537, MSA 554, 8325, Sanger MRSA, Sanger MSSA, Mu50, N315, and NCTC6571). Extensive recombination would result in *set* genes composed of segments of DNA with different evolutionary histories and conflicting phylogenetic topologies. However, we found that the topology representing the conserved regions of RD13 and each of the individual topologies constructed for genes *set4* to *set7* were consistent, suggesting that recombination had not occurred at a sufficiently high frequency to distort the phylogenetic signal.

The analysis suggested two models for the evolution of the RD13 locus (Fig. 3). One model requires a common ancestor which has only *set5* and *set7.* The present-day complement of *set* genes in each lineage would have arisen through multiple acquisition and deletion events. In this case, one would expect less variation in the form of synonymous (silent) nucleotide substitutions in the recently acquired *set* genes. However, our analysis indicated that the proportions of polymorphic synonymous sites in *set4* to *set7* are very similar, a result inconsistent with this model. Alternatively, the ancestral state could be represented by a complete complement of *set* genes in the Vr. In this model, extant states are explained by the loss of *set* genes in parallel in separate lineages of pathogenic *S. aureus* (Fig. 3). The occurrence of similar proportions of synonymous sites strongly supports this idea. Taken together, these analyses suggest that the evolution of the RD13 locus is explained by a model in which the loss of *set* genes has occurred several times independently in separate lineages of pathogenic *S. aureus* (Fig. 3). While the extent of horizontal gene transfer has been insufficient to mask the phylogenetic signal present, recombination may have contributed to chromosomal diversification at the RD13 locus. To investigate this possibility, the maximum χ^2 method was used to identify putative end points of past recombination. Multiple end points flanking small regions of the sequence throughout the RD13 locus were identified (data not shown). The largest of these regions was identified in the *res* gene based on a comparison of six strains (Fig. 3). Restricted allelic variation within the recombined segments of the *res* gene suggests that recombination has occurred recently.

Analysis of the level of selective constraint acting on *set* genes. Considering the high level of predicted homology among SET proteins produced by a single strain, it is conceivable that SETs may have some redundancy of function. Accordingly, we examined the possibility that one or more *set* genes may have become silent. Silent genes are free to accumulate synonymous (silent) mutations and corresponding amino acid replacements, as selective constraint is no longer acting to maintain protein function by limiting deleterious amino acid replacements. To determine if the level of selective constraint varies across RD13, we calculated p_N and p_S for subsets of 30 codons in a sliding window for the length of each *set* gene (Fig. 4). The difference, $p_N - p_S$, is a measure of the degree of selective constraint. The more negative the value, the

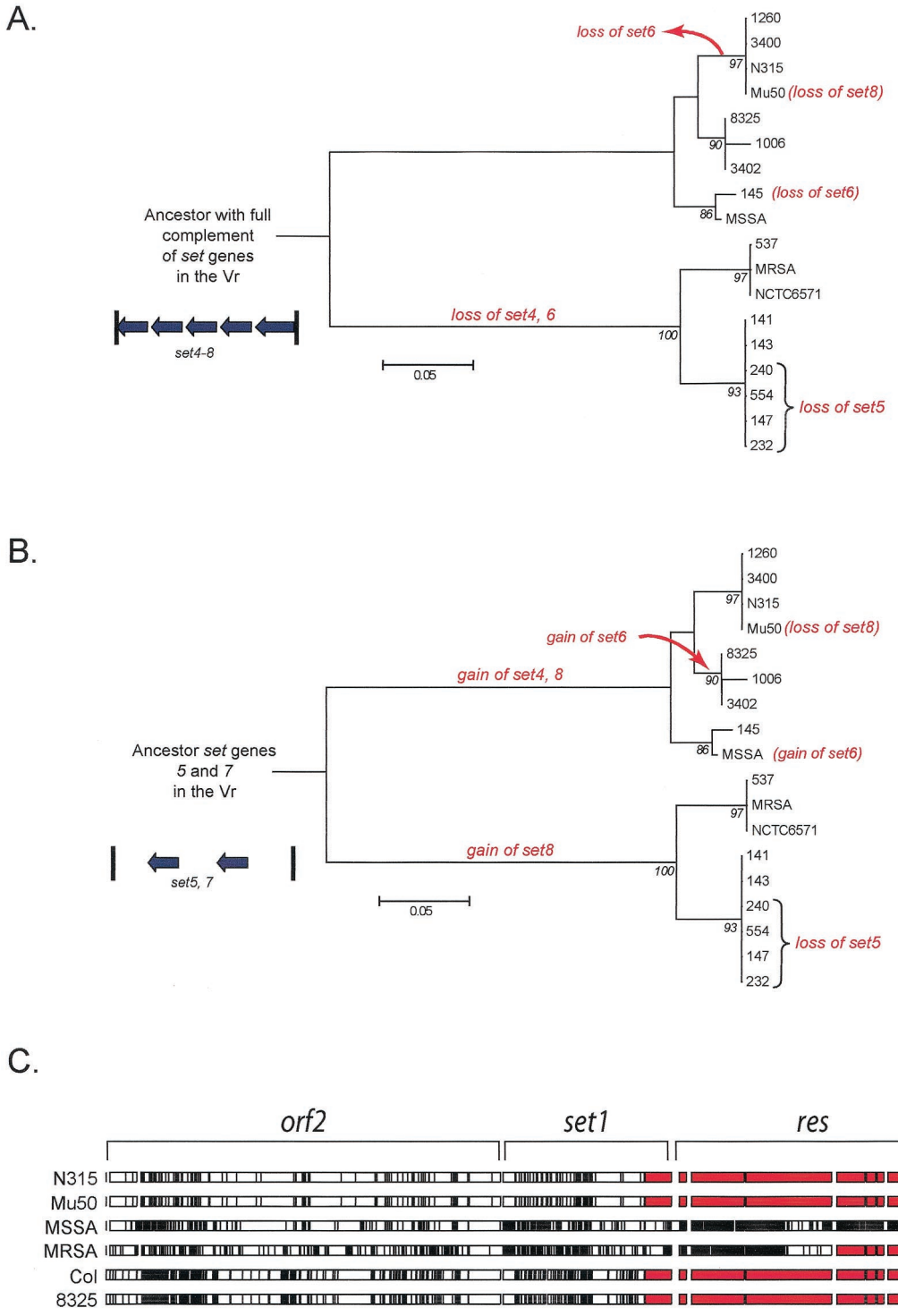

FIG. 3. Two models for the evolution of the RD13 chromosomal region and the Vr. The trees were constructed by the neighbor-joining algorithm based on the number of synonymous substitutions per synonymous site (d_S). A comparison of phylogenetic trees constructed from individual *set* gene nucleotide data and a concatenated sequence representing the conserved genes of RD13 indicated that the topologies were cognate. Thus, the extent of recombination present was insufficient to disrupt the underlying phylogenetic signal. The *set9* gene tree is shown. Bootstrap confidence limits are shown under the major nodes. The proposed ancestral state of the Vr is indicated at the hypothetical root. The gain and loss of *set* genes necessary to explain the extant state of each strain are indicated in red. (A) Evolution of RD13 and the Vr explained solely by the loss of *set* genes over time. This (preferred) model is supported by the observation that the proportion of silent mutations (synonymous substitution) within each *set* gene is the same. (B) Evolution of RD13 explained by both the gain and loss of *set* genes over time (alternative model). This model is not supported by an analysis of the proportion of polymorphic synonymous sites present in each *set* gene. While other models are possible, they require a larger number of evolutionary events (i.e., gain or loss), making them less likely. (C) Recombination has occurred within RD13. A Happlot analysis of the *orf2*, *set1*, and *res* genes in six representative strains is shown. Polymorphic nucleotide sites based upon pairwise comparisons are represented by vertical lines. Putative areas of recombination, regions of similar nucleotide sequence in different genetic backgrounds, are represented in red.

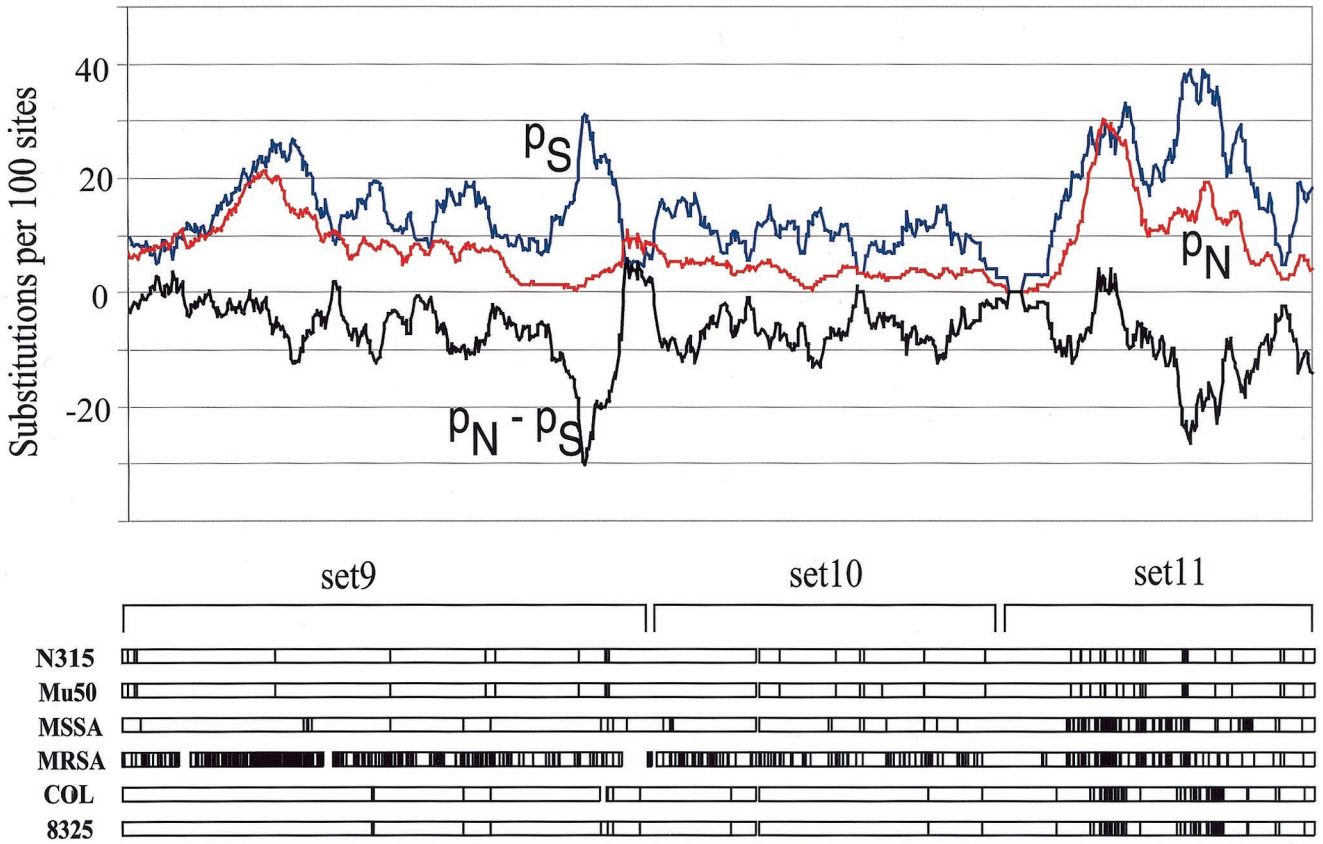

FIG. 4. Graphic display of the degree of selective constraint across a representative portion of RD13. The p_N and p_S nucleotide substitutions are indicated by the red and blue lines, accordingly. The black line represents the difference ($p_N - p_S$), which is a measure of the degree of selective constraint. The more negative the value, the less the contribution of amino acid replacements and the greater the contribution of synonymous-nucleotide substitutions. Overall, the difference ($p_N - p_S$) is consistently negative for each of the *set* genes, indicative of purifying selection. Polymorphic nucleotide sites based upon pairwise comparisons are represented by vertical lines.

less the contribution of amino acid replacements and the greater the contribution of synonymous nucleotide substitutions. A difference of zero indicates selectively neutral variation, where the per-site rates of synonymous and nonsynonymous substitutions are equal. A positive difference (i.e., amino acid replacements exceeding silent substitutions) suggests the action of diversifying (positive) selection. Overall, the $p_N - p_S$ difference is consistently negative for each of the *set* genes, a value indicative of purifying selection (Fig. 4). The *set9* gene did possess a single region 53 codons in length, stretching from position 17 to position 69, in which the value for $p_N - p_S$ was ≥ 0. In this region, the rate of substitution per 100 sites is expressed by the equations $d_N = 9.3 \pm 2.1$ and $d_S = 10.1 \pm 3.8$. The results suggest that if any of the *set* genes were silenced, it was a very recent event given that the number of corresponding amino acid replacements has not had sufficient time to accumulate substantially.

Expression of genes located in RD13 in *S. aureus* strain COL. To test the hypothesis that the genes in RD13 are transcribed, TaqMan assays were used to determine relative gene-specific mRNA levels present in mid-exponential and stationary-phase *S. aureus* cells (Fig. 5) by using the oligonucleotides listed in Table 4. No transcript was detected in either exponential- or stationary-phase cells for *set1*, hypothetical *orf5*, and the putative transposase (*tra)* genes, a result indicating that these three genes are not transcribed in vitro at these time points under the conditions studied. In contrast, all other genes examined were either constitutively expressed during the growth cycle (*res*, *mod*, and *set2*) or were up-regulated in the stationary phase of growth (*orf1* to *orf3*, *orf5*, *set3*, *set8*, *set9*, *set10*, and *set11*) (the cutoff was a 1.5-fold increase in the normalized transcript level).

Western blot analysis of recombinant SET proteins with sera from patients with invasive *S. aureus* infections. To determine if SET proteins are expressed during the course of human infection and stimulate a humoral immune response, Western immunoblot analysis of recombinant SET proteins was carried out with paired acute- and convalescent-phase serum samples from 19 patients with *S. aureus* bacteremia. Of the panel of 11 recombinant SET proteins, 6 were immunoreactive with sera obtained from at least one patient, indicating expression of these proteins during human infection (Table 5). SET2, SET4, SET5, SET8, SET9, and SET10 were reactive with sera from 13, 1, 4, 12, 18, and 3 of the 19 patients examined, respectively. SET1, SET3, SET6, SET7, and SET11 were not reactive with any of the 19 patient sera. Of the 11 recombinant proteins, only SET9 was reactive with any of the eight control sera tested ($n = 2$) (data not shown). A representative Western blot is shown in Fig. 6.

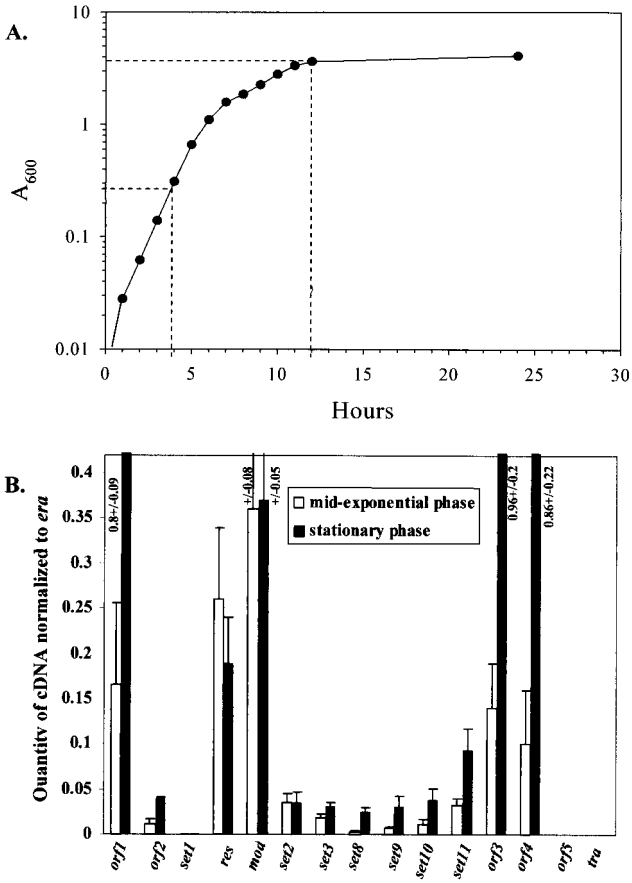

FIG. 5. (A) Growth curve of *S. aureus* strain COL. Total RNA was isolated from *S. aureus* cells grown in tryptic soy broth at 37°C to the A_{600} values and time points indicated by the dashed lines. (B) Relative quantities of RD13 reverse-transcribed mRNA normalized to the internal control *era*, determined by Taqman assays.

Biological activity of SET6. We next assessed whether a representative protein (SET10) encoded by a gene in the RD13 locus (Fig. 1) had superantigenic activity. In a standard assay, SET6 was unable to stimulate rabbit splenocyte or human PBMC proliferation whereas TSST-1 (positive control protein) stimulated proliferation of both rabbit splenocytes and human PBMCs (data not shown).

The pyrogenic activity of purified SET6 and its ability to enhance endotoxic shock also were examined. SET10 did not cause fever in rabbits at a dose of 10 μg/kg and did not enhance the lethality of endotoxin administered to rabbits. TSST-1 administered at the same dose was pyrogenic and enhanced the lethality of endotoxin (data not shown).

DISCUSSION

Several comparative genomic studies have highlighted the extensive variation that exists within natural populations of some pathogenic bacteria, but the molecular mechanisms of genome diversification are still unclear. We identified extensive variation in RD13, a region of the chromosome that is common to all strains of *S. aureus* examined. This variation included the sequence divergence of genes common to all strains and differences in gene content, as indicated by RD13

TABLE 4. TaqMan assay oligonucleotide primers and probes used to quantitate cDNA

Gene	Forward primer (F), reverse primer (R), and probe (P) (5′–3′)
orf1	F-GGCAATGGTAGGTGTTTTAGCAA, R-ACGAGTCCATTTTGAGAATAAACTTTC, P-TGCTTGATTACCATATCCAACAACGCCA
orf2	F-GGAAAACACACCTGCATATCATAAA, R-ACCTAATGGATAAAACCAGGTAATCG, P-ATCCGTTAGTATCGTAGCCGACAATTTTATCACCAT
set1	F-TGAAAGATGGTGGGTTCTACACA, R-TTTTTTCTATATTTCTGCCATCAATAACA, P-CCCATACGGTGTGTTTGTAACTTTTTATTCAATTCAA
res	F-AAAAGAAAAATGTGCCAGAATTGAG, R-CTGTAAGATCCCCTAACTGCTTCTCT, P-CCATTCGCCTTCAAACCCTGGGAA
mod	F-TGTTAGGTGATGCATATGAATTCCTAA, R-ATAGAACTCGCCTGCTTTTTTACC, P-GTCGCCGCAAAGCGCCCA
set2	F-CAGAAGTTCATTCAGGTCATGCA, R-CCATAGTCTTTCCAGTGTAGTATCGGTATA, P-AACAAAATCAAAAGTCAGTAAATAACATGACAAGGAAGC
set3	F-GCATTAGGAATATTAACTACAGGTGTGTTT, R-TGCGTTGTGTCTCATCAAGTTCT, P-TCGCGTGACCAGTTTGACTTTCTCTGCTG
set8	F-ACGGTGGAAAGTACACGTTTGA, R-ACTTCGATGTTTTTAATTTGTTCACTATTA A, P-ACATCTGCCATGCGATTTTCTTGTAATTTTTTG
set9	F-AAATGAGAACAATTGCTAAAACCAGTT, R-ACCGATTGCGTCGTTACTGTAA, P-AGCACTAGGGCTTTTTAACAACAGGCGCA
set10	F-TGAAGGTGCAAAGTACTCTATTGGG, R-CAGTATGATCTTCTTTAATAACTCTTGCTTCT, P-TCGACAGCTTTATCGTTTGCACTCGTGAT
set11	F-GGGAATGTTAGCAACAGGTGTAATT, R-TGTTTTAACTCTGATTCACTTTGTTGTTT, P-CATCGAATGTACAATCAGTACAAGCGAAAGCA
orf3	F-GGCATCAGGCACAGAAATTAAAT, R-CCAAAGCCTTTAGGTTCATCATACA, P-TGAGCCCGTTTCATTAGAGACATTTGCAG
orf4	F-TTTAGTGTATATCATTGGCGTTGCTA, R-CAATACTGGCAATGTGCCTTGT, P-TTATGGTGGCCTTTCAAGCAATATTAGTATCTTTTTCA
orf5	F-GATTTGTTTTTTTGTGGTTATTTTTTAATT, R-CACAAGGAGTGATTATCATGGTACCA, P- TGCAATCGCAAAATTCAAACCATTTAAGAACA
tra	F-GATTTGTTTTTTTGTGGTTATTTTTTAATT, R-CACAAGGAGTGATTATCATGGTACCA, P-TCGATAAAAACAACACCAAAATACTGGACCCTTCA
era	F-TTGAAAGAGATTCGCAAAAAGGA, R-CACGTCTCGCACGTTTTTCC, P-TTCTTTTAACTTTTTACCGCCTTTTCCAATGACAA

TABLE 5. Immunoreactivity of recombinant SET proteins with human patient sera by Western blot analysis

Patient serum sample[a]	Reactivity of recombinant SET protein[b]										
	SET1	SET2	SET3	SET4	SET5	SET6	SET7	SET8	SET9	SET10	SET11
1a		+			+			+	+	+	
1c		+			+			+	+	+	
2a		+						+	+		
2c		+						+	+		
3a									+		
3c					+			+	+		
4a									+		
4c		+						+	+		
5a									+		
5c									+		
6a								+	+		
6c								+	+		
7a											
7c									+		
8a		+		+				+	+	+	
8c		+		+	+			+	+	+	
9a								+	+		
9c		+			+			+	+		
10a									+		
10c		+							+		
91a									+		
91c		+							+		
92a		+						+	+		
92c		+						+	+		
93a								+	+		
93c		+						+	+		
94a											
94c									+		
96a									+		
96c									+		
97a		+						+	+		
97c		+									
98a		+							+		
98c		+						+	+		
99a		+						+	+		
99c		+						+	+		
100a		+								+	
100c		+								+	

[a] Human serum samples were obtained from individuals with invasive episodes of *S. aureus* infection in Finland in 1999 and 2000. Patients had infections with or without deep infection foci, including abscesses, spondylitis, amnionitis, foreign body infections, pneumonia, and purulent arthritis. Sera were obtained 2 to 7 days after the first positive blood culture (a, acute phase) and 20 to 30 days postinfection (c, convalescent phase).

[b] +, positive reactivity.

PCR-RFLP, size, and DNA sequence analyses. The seven *S. aureus* strains whose genomes have been sequenced had an RD13 central Vr of 7, 9, 10, or 11 kb corresponding to 6, 8, 9, or 10 *set* genes, respectively (Fig. 1). Most of the 63 strains examined by PCR had an RD13 with one of these Vr size variants. However, four strains had a Vr of 8 kb, consistent with the presence of seven *set* genes in the Vrs in these strains. This inference was confirmed by DNA sequencing.

The presence of a large and variable number of *set* genes and the observed sequence divergence in the RD13 locus suggested that molecular evolutionary genetic analysis would provide insight into the processes shaping variation in this region. To investigate the evolution of RD13, we adopted a strategy in which separate phylogenies were constructed to represent the conserved architecture of the locus and the variable *set* genes. The analysis suggested two models of evolution for RD13 (Fig. 3). One model requires a common ancestor which possesses only *set5* and *set7*. In this situation, multiple gene acquisition and gene loss events would have led to the present-day complement of *set* genes in each lineage. However, based on a similar proportion of synonymous (silent) nucleotide substitutions in all *set* genes in the Vr, the most likely scenario is that an ancestral strain with a full complement of *set* genes in the Vr underwent multiple independent losses of *set* genes in parallel in distinct lineages. In each case, the *set* gene losses were confined to the *set4* to *set8* genes. It is unlikely that there is a selective disadvantage to possessing one or more of the *set4* to *set8* genes, as two strains (8325 and MSSA) have a copy of each gene in their extant states. In addition, analysis of nucleotide substitution in the *set4* to *set8* genes did not reveal a marked decrease in functional constraint that may accompany a deleterious allele.

The presence of RD13 in all strains representing the major clonal lineages within the species and the existence of a G+C content which is equivalent to that of the *S. aureus* genome both suggest that RD13 is an ancient feature of the *S. aureus* chromosome. DNA sequence analysis indicated that horizontal gene transfer and recombination have contributed to the diversification of RD13 in different strains (Fig. 3). The presence of a transposase gene in this region supports this theory, although this gene was not expressed in strain COL under the conditions studied and has been lost altogether by the recently

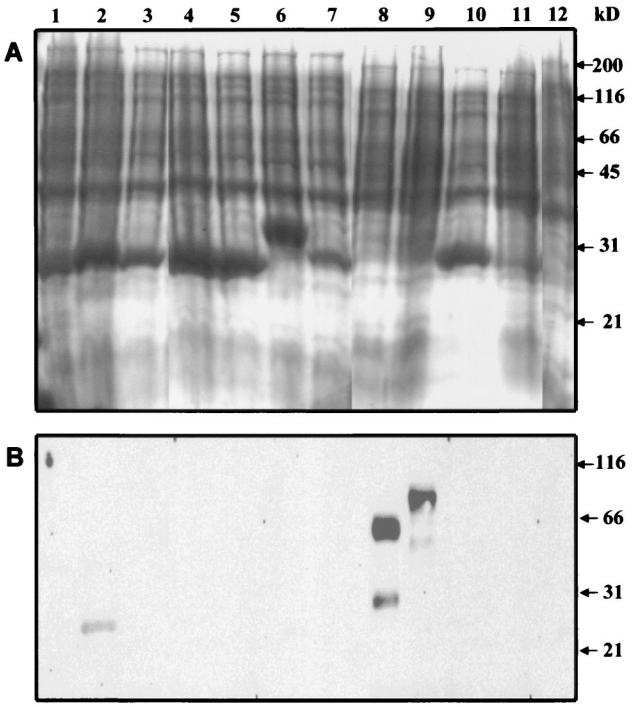

FIG. 6. (A) Immunogenicity of recombinant SET proteins with human patient sera. SDS-PAGE gel of lysates of *E. coli* expressing recombinant SET protein variants. Lanes 1 to 11, SET1 to SET11, respectively; lane 12, lysate of *E. coli* containing the expression vector only. Note that SET8 and SET9 resolve at approximately twice their expected molecular weight, thus indicating a dimeric form of the protein (see Western blot analysis in panel B). (B) Representative Western blot analysis of recombinant SET proteins with convalescent-phase serum from human patient 2. Lanes 1 to 11, SET1 to SET11, respectively.

sequenced strain MW2. The presence of many related but not identical *set* genes is suggestive of distant gene amplification events followed by sequence divergence which resulted in a common *S. aureus* ancestor with a full complement of *set* genes in RD13 (Fig. 3).

Our data show that the restriction-modification system genes in RD13 are expressed in vitro. Interestingly, restriction-modification system genes are found in another large variable region in the *S. aureus* chromosome adjacent to a cluster of six genes encoding a family of serine proteases (6). The significance of their close proximity has yet to be established, but it is possible that restriction-modification systems may have played a role in the evolution or regulation of some paralogous gene families in the *S. aureus* chromosome.

DNA sequence analysis of the ratio of synonymous nucleotide substitutions in *set* genes to amino acid substitutions in their products suggested that the functionality of *set* genes has been maintained by purifying selection (Fig. 4). Gene expression assays indicated that only one of seven *set* genes (*set1*) in strain COL was not expressed in vitro under the conditions tested. The *set* genes were expressed in a growth-phase-dependent manner (Fig. 5) with the highest transcription levels detected during the stationary phase of growth. In this regard, the *set* genes are similar to many staphylococcal exotoxin genes associated with virulence which are typically up-regulated postexponentially. Of note, the *set* genes were expressed concurrently, suggesting that multiple SET proteins may participate in host-pathogen interactions simultaneously. The identification of putative promoters upstream of several *set* genes (data not shown) suggests that the *set* genes may be encoded by multiple transcripts. The cotranscription of six *set* genes also indicates that a mechanism of phase variation mediated by gene cassette switching does not occur at this locus. Considering the evident selective constraint on SET proteins, it is likely that *set* gene deletions are a relatively rare occurrence. Of note, no RD13 gene deletions were observed during continuous in vitro passage of a clinical strain of *S. aureus* for a 6-week period, thus indicating the stability of this chromosomal region during growth in vitro (17).

The evolutionary basis for maintenance of production of many related SET proteins by *S. aureus* is unclear. However, the data indicate that multiple SET proteins are expressed during human infection and induce a humoral immune response. Although all patient sera examined contained antibodies to one or more recombinant SET proteins from strain COL or 8325, not all SET variants were immunoreactive. The possibility that different strains may produce antigenically distinct SET variants is consistent with the variation in the *set* gene content and DNA sequence identified in different strains. It is conceivable that SET allelic variation may represent a mechanism of immune avoidance employed by *S. aureus* which may encounter hosts that have already been exposed to SET proteins during previous staphylococcal infections. Alternatively, the variation in immunoreactivity in human patient sera may be explained by the possibility that some SET proteins did not elicit a strong immune response or are poorly expressed in some strains. The SET protein family may have evolved as a result of the occurrence of host receptors which are polymorphic between individuals, such that the production of many SET proteins may increase the chances of effective interaction with such receptors. However, it is also conceivable that although the SET proteins have high sequence homology and similar predicted structures, each protein has a related but distinct function in host-pathogen interactions.

We examined a SET variant (SET6) for superantigen activity. In spite of the presence of the two superantigen consensus domains, SET6 did not have the classical properties of superantigens such as superantigenicity, pyrogenicity, or enhancement of endotoxic shock (data not shown). This finding is consistent with that of a recent study (1) of a different SET variant (SET3), which indicates that SET proteins may have biological activities that are distinct from those of superantigens. The study by Arcus et al. (1) suggests that SET proteins have some of the characteristics of DNA-binding proteins, and further analysis will help define a clearer role for SET proteins in pathogenesis.

In this study, we explored the molecular evolution of a large chromosomal region common to all strains of *S. aureus* which encodes a family of proteins putatively involved in pathogenesis. We found extensive genetic diversity and variation in the *set* gene content at the chromosomal region RD13 within the *S. aureus* species, and we proposed a likely evolutionary scenario for the locus in contemporary pathogenic *S. aureus* strains that involves multiple episodes of *set* gene deletion in different clonal lineages in parallel. We also found that horizontal gene transfer and recombination have contributed to the diversification of this chromosomal locus. Functional constraint is acting to maintain the expression of *set* genes, and multiple SET proteins are expressed during human infections, suggesting a possible role in host-pathogen interactions. Taken together, these findings represent new insights bearing on the evolution and diversification of the genome of pathogenic *S. aureus*.

ACKNOWLEDGMENTS

We thank J. Voyich for technical assistance and M. Otto for critical review of the manuscript.

REFERENCES

1. **Arcus, V. L., R. Langley, T. Proft, J. D. Fraser, and E. N. Baker.** 2002. The three-dimensional structure of a superantigen-like protein, SET3, from a pathogenicity island of the *Staphylococcus aureus* genome. J. Biol. Chem. **277:**32274–32281.
2. **Baba, T., F. Takeuchi, M. Kuroda, H. Yuzawa, K. Aoki, A. Oguchi, Y. Nagai, N. Iwama, K. Asano, T. Naimi, H. Kuroda, L. Cui, K. Yamamoto, and K. Hiramatsu.** 2002. Genome and virulence determinants of high virulence community-acquired MRSA. Lancet **359:**1819–1827.

2a. **Bradford, M. M.** 1976. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Anal. Biochem. **72:**248–254.

3. **Chaussee, M. S., R. O. Watson, J. C. Smoot, and J. M. Musser.** 2001. Identification of Rgg-regulated exoproteins of *Streptococcus pyogenes.* Infect. Immun. **69:**822–831.
4. **Dinges, M. M., P. M. Orwin, and P. M. Schlievert.** 2000. Exotoxins of *Staphylococcus aureus.* Clin. Microbiol. Rev. **13:**16–34.
5. **Fitzgerald, J. R., and J. M. Musser.** 2001. Evolutionary genomics of bacterial pathogens. Trends Microbiol. **9:**547–553.
6. **Fitzgerald, J. R., D. E. Sturdevant, S. M. Mackie, S. R. Gill, and J. M. Musser.** 2001. Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. Proc. Natl. Acad. Sci. USA **98:**8821–8826.
7. **Foster, T. J., and M. Hook.** 1998. Surface protein adhesins of *Staphylococcus aureus*. Trends Microbiol. **6:**484–488.
8. **Hunter, P. R.** 1990. Reproducibility and indices of discriminatory power of microbial typing methods. J. Clin. Microbiol. **28:**1903–1905.
9. **Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei.** 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17:**1244–1245.
10. **Kuroda, M., T. Ohta, I. Uchiyama, T. Baba, H. Yuzawa, I. Kobayashi, L. Cui, A. Oguchi, K. Aoki, Y. Nagai, J. Lian, T. Ito, M. Kanamori, H. Matsumaru, A. Maruyama, H. Murakami, A. Hosoyama, Y. Mizutani-Ui, N. K. Taka-**

hashi, T. Sawano, R. Inoue, C. Kaito, K. Sekimizu, H. Hirakawa, S. Kuhara, S. Goto, J. Yabuzaki, M. Kanehisa, A. Yamashita, K. Oshima, K. Furuya, C. Yoshino, T. Shiba, M. Hattori, N. Ogasawara, H. Hayashi, and K. Hiramatsu. 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus.* Lancet **357:**1225–1240.

11. **Marr, J. C., J. D. Lyon, J. R. Roberson, M. Lupher, W. C. Davis, and G. A. Bohach.** 1993. Characterization of novel type C staphylococcal enterotoxins: biological and evolutionary implications. Infect. Immun. **61:**4254–4262.
12. **Musser, J. M., and R. K. Selander.** 1990. Genetic analysis of natural populations of *Staphylococcus aureus*, p. 59–67. *In* R. P. Novick (ed.), Molecular biology of the staphylocci. VCH Publishers, Inc., New York, N.Y.
13. **Nei, M., and T. Gojobori.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:**418–426.
14. **Papageorgiou, A. C., and K. R. Acharya.** 2000. Microbial superantigens: from structure to function. Trends Microbiol. **8:**369–375.
15. **Reid, S. D., R. K. Selander, and T. S. Whittam.** 1999. Sequence diversity of flagellin (*fliC*) alleles in pathogenic *Escherichia coli.* J. Bacteriol. **181:**153–160.
16. **Smith, J. M.** 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **34:**126–129.
17. **Somerville, G. A., S. B. Beres, J. R. Fitzgerald, F. R. DeLeo, R. L. Cole, J. S. Hoff, and J. M. Musser.** 2002. In vitro serial passage of *Staphylococcus aureus*: changes in physiology, virulence factor production, and *agr* nucleotide sequence. J. Bacteriol. **184:**1430–1437.
18. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
19. **Williams, R. J., J. M. Ward, B. Henderson, S. Poole, B. P. O'Hara, M. Wilson, and S. P. Nair.** 2000. Identification of a novel gene cluster encoding staphylococcal exotoxin-like proteins: characterization of the prototypic gene and its protein product, SET1. Infect. Immun. **68:**4407–4415.

Editor: D. L. Burns